Project 1: Exploratory Data Analysis

Due October 8th at 11:59 pm on Canvas

Project Description

This project is a collaboration with Dr. Lauren Micalizzi from the Department Behavioral and Social Sciences. Exposure to smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) are two of the most ubiquitous and hazardous of children's environmental exposures. Seven to 15% of infants born per year are exposed to SDP and more than 25% of children are exposed to household ETS. SDP alone imposes **a** $4 billion annual economic burden due to health-care costs. Early smoke-exposure increases rates of externalizing behaviors in children, including Attention-Deficit/Hyperactivity Disorder, and rates of substance use problems, all of which have clear public health implications. Early smoke exposure is also linked to self-regulation problems (i.e., the maintenance of physiological, emotional, behavioral, and cognitive control).

The purpose of Dr. Micalizzi's research is to examine the association between smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) exposure and self-regulation, externalizing behavior, and substance use. The women in this study were recruited from a previous study on smoke avoidance intervention to reduce low-income women's (N=738) smoking and ETS exposure during pregnancy and children's exposure to ETS in the immediate postpartum period. A subset of adolescents (N=100) and their mothers are randomly selected for recruitment into this study. You will have access to the baseline data; two longitudinal follow up assessments occur at 6- and 12-months post-baseline.

Your work on this dataset will is related to Aim 1 of 3 primary study aims:
AIM 1: Examine effects of SDP/ETS on adolescent self-regulation, substance use, and externalizing.
AIM 2: Explore links between self-regulation at baseline and substance and externalizing at 6- and 12-month follow-ups.
AIM3: Identify self-regulation problems that mediate the link between SDP/ETS exposure and level of, and change in, SU and EXT severity over time.

Dr. Micalizzi will be coming to class on TBD to share more background about this project. The data for this project is available on Canvas along with a corresponding codebook. Be sure to read the codebook carefully before Dr. Micalizzi's presentation to be ready to ask questions.

Data Privacy: This data can only be shared with the instructor, students, and TA of this class. In particular, you should not put the data in your group's github repository or anywhere else that people could potentially access it. Your report and code itself can be public as long as it does not contain any raw or identifiable data and instead presents results at a summary level. For example, showing a scatter plot or summary table is okay, but printing the head of the data is

not. Your reports will be shared with Dr. Micalizzi. If she uses any analysis from your reports in future publications or presentations, she will contact you about acknowledgment.

Report and Github Repository

Your goal will be to conduct an exploratory data analysis of this data driven by the research goals of the study. As part of your analysis, you should explore the univariate and bivariate relationships in the data as well as any missing data. You should think about what questions we can ask based on our goals and create tables and figures that are informative to the reader. Given the amount of variables available in this data, you should think about how you will select the information you include. Your exploratory analysis can contain regression analysis if helpful.

You will submit your analysis as a pdf report on Canvas. There is a 12-page limit for the report, excluding a code appendix. Additionally, your report should be available in your class github repository with your code. This may contain code beyond your R markdown file since you may not choose to include all your analysis in your report. Your repository should have a short readme with a description of the project and the files.

Rubric

Your project will be graded according to the following criteria. See the attached rubric for more information.

- Writing, Figures, and Tables (5 pts)
- Reproducible Code (5 pts)
- Exploratory Analysis (10 pts)
- Overall Impact and Contribution (5 pts)